PROBABILIDAD Y ESTADÍSTICA II

TEMA 4. REDES DE COLAS

CONTENIDOS

1. Introducción a las redes de colas

De todos los elementos básicos que componen un sistema de colas, tan sólo nos queda discutir sobre el quinto elemento: el número de etapas de servicio.

Hasta ahora los clientes demandaban del sistema una sola operación de servicio. Por eso los sistemas eran de un solo nodo, donde quizá podía haber varios servidores idénticos paralelos.

Ahora nos interesan sistemas con múltiples nodos en los que el cliente requiere servicio en más de uno.

Así, los clientes pueden entrar al sistema por varios nodos, encolarse para ser servidos y salir de un nodo dado para entrar en otro y recibir servicio adicional o para abandonar el sistema definitivamente. No todos los clientes entran y salen del sistema por los mismos nodos necesariamente, o siguen el mismo camino una vez en el sistema. Los clientes pueden regresar a nodos previamente visitados, saltarse algunos e incluso escoger permanecer en el sistema para siempre.

Las redes de colas son un conjunto de nodos interrelacionados que funcionan de forma asíncrona (entradas y salidas de clientes no tienen que estar sincronizadas) y concurrente (simultáneamente).

La mayoría de los sistemas informáticos son sistemas con múltiples nodos. Pueden tener terminales on-line, líneas de comunicación, impresoras, controladores de comunicación y el propio ordenador, por ejemplo.

Las redes de colas se clasifican en dos grupos. En las redes abiertas los clientes pueden entrar y salir del sistema. En las redes cerradas no entran nuevos clientes y los existentes nunca salen, es decir, el número de clientes es constante a lo largo del tiempo, como en el modelo de reparación de máquinas, que es un ejemplo de red cerrada.

La estructura topológica de la red es importante porque describe las transiciones admisibles entre nodos (no deben confundirse con las transiciones entre los estados del sistema). También deben describirse los caminos recorridos por los clientes, así como los procesos estocásticos que configuran el flujo (estocástico) que transcurre por la red.

2. Redes de colas abiertas

Teorema de Burke. El proceso de salidas de clientes de un sistema *M/M/c* estable ($\lambda/c\mu < 1$) con tasa de llegadas λ es un proceso de Poisson de tasa λ.

(Demostración: Gross y Harris (1998) p.168 o Kleinrock (1975), p.148)

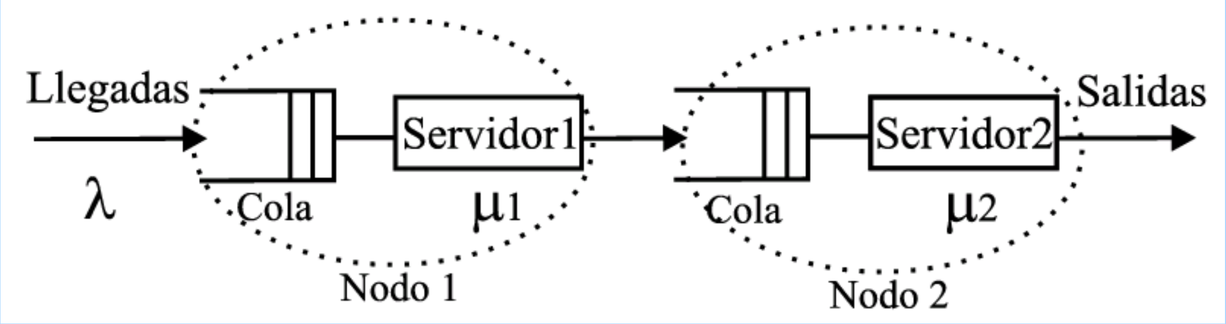
La distribución del tiempo entre salidas consecutivas de un *M/M/c* es idéntica a la distribución del tiempo entre llegadas, es decir, exponencial de parámetro λ.

Así, la distribución de las salidas es como la de las llegadas y no se ve afectada por el mecanismo de servicio exponencial. Se puede demostrar además que los tiempos entre salidas consecutivas son independientes entre si.

2.1 Sistema en tándem

El primer caso a analizar es un sistema tándem, también denominado secuencial o en serie.

Consideramos un sistema con dos procesadores (servidores) en el que los clientes llegan con tasa λ según un proceso de Poisson. Después de ser servidos por el procesador 1 se unen a la cola del procesador 2. Suponemos que ambas colas disponen de capacidad ilimitada. Cada procesador sirve en tiempo exponencial con tasa μ_i, i=1,2.



El estado del sistema será un par (n,m) que indica que hay n clientes en el nodo 1 y m en el nodo 2. Las ecuaciones de equilibrio son

Estado	Tasa entrada	= Tasa salida
$(0,0)$	$\mu_2\pi_{0,1}$	$= \lambda\pi_{0,0}$
$(n,0), n>0$	$\lambda\pi_{n-1,0} + \mu_2\pi_{n,1}$	$= (\lambda+\mu_1)\pi_{n,0}$
$(0,m), m>0$	$\mu_1\pi_{1,m-1} + \mu_2\pi_{0,m+1}$	$= (\lambda+\mu_2)\pi_{0,m}$
$(n,m), nm>0$	$\lambda\pi_{n-1,m} + \mu_1\pi_{n+1,m-1} + \mu_2\pi_{n,m+1}$	$= (\lambda+\mu_1+\mu_2)\pi_{n,m}$

junto con la ecuación usual $\sum_{n,m} \pi_{n,m} = 1$.

Sea π_n^1 la probabilidad de que haya *n* clientes en el nodo 1 y π_m^2 la probabilidad de que haya *m* clientes en el nodo 2. La situación del nodo 1 es la de un modelo *M/M/1*. Por el teorema de Burke, la situación del nodo 2 corresponde también a un *M/M/1*. Luego,

$$\pi_n^1 = \left(\frac{\lambda}{\mu_1}\right)^n \left(1 - \frac{\lambda}{\mu_1}\right); \quad \pi_m^2 = \left(\frac{\lambda}{\mu_2}\right)^m \left(1 - \frac{\lambda}{\mu_2}\right).$$

Ahora, si el número de clientes en el nodo 1 y en el 2 fueran variables aleatorias independientes, se tendría que $\pi_{n,m} = \pi_n^1 \, \pi_m^2$.

Veamos que ésta es precisamente la solución del sistema en equilibrio. Para ello sólo hay que comprobar que satisface todas las ecuaciones, ya que sabemos que la solución es única. Para la primera ecuación, hay que verificar

$$\mu_2 \left(1 - \frac{\lambda}{\mu_1}\right) \left(\frac{\lambda}{\mu_2}\right) \left(1 - \frac{\lambda}{\mu_2}\right) = \lambda \left(1 - \frac{\lambda}{\mu_1}\right) \left(1 - \frac{\lambda}{\mu_2}\right),$$

que es inmediato. Con el resto de ecuaciones se procedería de forma análoga.

Por tanto, $\pi_{n,m} = \pi_n^1 \, \pi_m^2$ es la solución estacionaria y el número de clientes en el nodo 1 es independiente del número de clientes en el nodo 2. Esto no implica que los tiempos de espera de un cliente en las dos colas sean independientes, como puede demostrarse. Sin embargo, los tiempos totales (añadiendo el tiempo en el servidor) sí lo son.

Se tiene que

$$L = \sum_{n,m}(n+m)\pi_{n,m} = \sum_{n} n\pi_n^1 + \sum_{m} m\pi_m^2 = \frac{\lambda}{\mu_1 - \lambda} + \frac{\lambda}{\mu_2 - \lambda}.$$

Y por la fórmula de Little

$$W = \frac{L}{\lambda} = \frac{1}{\mu_1 - \lambda} + \frac{1}{\mu_2 - \lambda}.$$

Además,

$$W_q = W - \left(\frac{1}{\mu_1} + \frac{1}{\mu_2}\right).$$

2.2 Redes de Jackson abiertas

Los resultados precedentes con una distribución estacionaria tan útil se generalizan en gran medida a las redes de Jackson:

(Teorema de Jackson) *Sea una red de colas, denominada red de Jackson, con R nodos que satisfacen las siguientes condiciones:*

1. *Cada nodo i consiste en c_i servidores idénticos, cada uno con tiempo de servicio exponencial de tasa μ_i.*
2. *La llegada de clientes al nodo i desde fuera del sistema es según un proceso de Poisson de tasa λ_i. (También pueden llegar clientes al nodo i desde otros nodos de dentro de la red).*
3. *Una vez servido en el nodo i, el cliente pasa (instantáneamente) al nodo j ($j = 1, 2, ..., R$) con probabilidad p_{ij} o abandona la red con probabilidad $1 - \sum_{j=1}^{R} p_{ij}$.*

Entonces, para cada nodo i $(i = 1, 2, ..., R)$, *la tasa media de llegadas totales al nodo* i, Λ_i, *es*

$$\Lambda_i = \lambda_i + \sum_{j=1}^{R} \Lambda_j p_{ji}.$$

Además, si $\pi(n_1, ..., n_R)$ *denota la probabilidad estacionaria de que haya* n_1 *clientes en el nodo* $1, ..., n_R$ *clientes en el nodo* R, *y* $\Lambda_i < c_i \mu_i$ *para todo* i, *entonces*

$$\pi(n_1, ..., n_R) = \pi_1(n_1) \cdots \pi_R(n_R),$$

donde $\pi_i(n_i)$ *es la probabilidad estacionaria de que haya* n_i *clientes en el nodo* i *si éste se trata como un sistema* $M/M/c_i$ *con tasa media de llegadas* Λ_i *y tasa media de servicio* μ_i *en cada servidor. Más aún, cada nodo* i *se comporta como si fuese un sistema* $M/M/c_i$ *independiente, con llegadas de Poisson de tasa* Λ_i.

Las ecuaciones para los Λ_i son intuitivas porque λ_i es la tasa de llegadas al nodo *i* desde fuera del sistema, y como Λ_j es la tasa a la que los clientes salen del nodo *j* (la tasa de entrada debe ser igual a la de salida), $\Lambda_j\, p_{ji}$ es la tasa de llegada a *i* de aquellos que vienen de *j*.

Nótese que el teorema de Burke permitía sólo conexiones hacia delante, sin realimentación, ya que podía destruir la naturaleza poissoniana del caudal de salida realimentado.

Por eso, si hay realimentación, el proceso de llegadas totales a un nodo (exteriores más realimentadas) no será de Poisson. Asombrosamente, el teorema de Jackson indica que incluso las redes con realimentación son tales que los nodos se comportan "como si" fueran alimentados totalmente por llegadas de Poisson, aunque en realidad no sea así.

En Λ_i estamos sumando las llegadas (de Poisson) desde fuera del sistema y las llegadas (no necesariamente de Poisson) desde todos los nodos internos.

Las probabilidades estacionarias en cada nodo son las de un modelo *M/M/*c_i, incluso aunque el modelo no sea un *M/M/*c_i. Los estados n_i de los nodos individuales son v.a. independientes.

La condición $\Lambda_i < c_i \mu_i$ para todo i es necesaria para que todos los nodos de la red representen cadenas de Markov ergódicas. Esta formulación tan general permite el caso en que $p_{ii} \geq 0$. La tasa de salida (externa) del sistema desde el nodo i es

$$\Lambda_i \left(1 - \sum_{j=1}^{R} p_{ij}\right)$$

Por la forma producto de la probabilidad estacionaria, resulta que el número medio de clientes en el sistema, L, es la suma del número medio de clientes en cada nodo L_i, como vimos en las colas en serie. A partir de L, podemos calcular W como

$$W = \frac{L}{\lambda}, \quad \text{donde } \lambda = \sum_{i=1}^{R} \lambda_i.$$

Sobre las distribuciones de los tiempos de espera no se puede decir mucho. El hecho de que los nodos se comporten como si fueran modelos $M/M/c_i$ nos puede hacer pensar que podríamos usar las distribuciones de los tiempos de espera de esos modelos. Sin embargo, esto no es necesariamente cierto en redes de Jackson, donde se permite la realimentación.

Los sistemas tándem son redes de Jackson. En el caso más general de colas en serie con R nodos en lugar de 2, en el teorema de Jackson se tiene que $\lambda_i = \lambda$ para $i = 1$ y $\lambda_i = 0$ en el resto, y además $p_{i,i+1} = 1$ para $i = 1,2,\ldots,R-1$, $p_{Rj} = 0$ para $j = 1,\ldots,R$, y son también redes de Jackson.

Hemos supuesto capacidad infinita en los nodos. Analizar redes de colas cuando hay límites en la capacidad de algún nodo es más complejo.

Puede que haya un efecto de bloqueo; esto es, si un cliente ha terminado su servicio en el nodo i y quiere dirigirse a un nodo j que está al máximo de su capacidad, entonces debe esperar en el nodo i hasta que haya sitio en el nodo j y el sistema se bloquea. Las llegadas al nodo i se rechazan.

Otra posibilidad es que ese cliente rebose el nodo j y deba irse inmediatamente a otro nodo en su lugar. Una última opción es que ese cliente sea rechazado y tenga que abandonar el sistema.

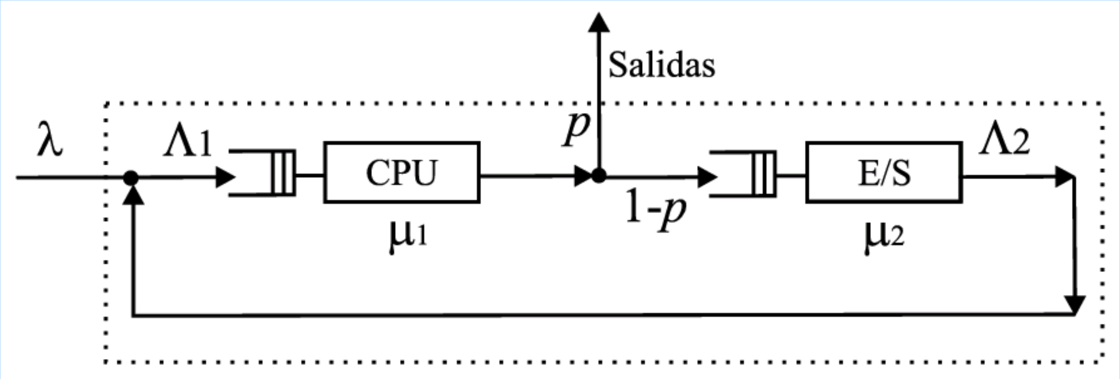
2.3 Aplicación: Multiprogramación

En un sistema de multiprogramación se almacenan en memoria principal varios programas simultáneamente. Cada programa es una secuencia de instrucciones de *CPU* y de entrada/salida (E/S).

Mientras el dispositivo de E/S está procesando alguna entrada o salida de un programa cuya terminación es requisito para poder seguir con más instrucciones de *CPU*, la *CPU* procesa otro programa.

Por tanto, la ejecución de un programa en este sistema sigue un movimiento cíclico entre la *CPU* y el dispositivo de *E/S*, hasta completar la ejecución (y salir del sistema). La red de colas asociada es cíclica con dos nodos.

Suponemos que cuando termina un servicio en la *CPU*, el programa abandona el sistema con probabilidad p o se encola para ser servido en la *E/S* con probabilidad $1-p$.



Las colas son de capacidad infinita. Por el teorema de Jackson, podemos calcular Λ_1 y Λ_2, las tasas de llegadas a los nodos de *CPU* y *E/S*, respectivamente:

$$\Lambda_1 = \lambda + \Lambda_2 = \lambda + (1-p)\Lambda_1 \Rightarrow \Lambda_1 = \frac{\lambda}{p}$$

$$\Lambda_2 = (1-p)\Lambda_1 = (1-p)\frac{\lambda}{p}$$

Las utilizaciones de los dos servidores son:

$$\rho_1 = \frac{\Lambda_1}{\mu_1} = \frac{\lambda}{p\mu_1}$$
$$\rho_2 = \frac{\Lambda_2}{\mu_2} = (1-p)\frac{\lambda}{p\mu_2}$$

La productividad media o paso a través del sistema es $p\Lambda_1 = \lambda$ trabajos por unidad de tiempo, lo que es cierto si no se pierden trabajos.

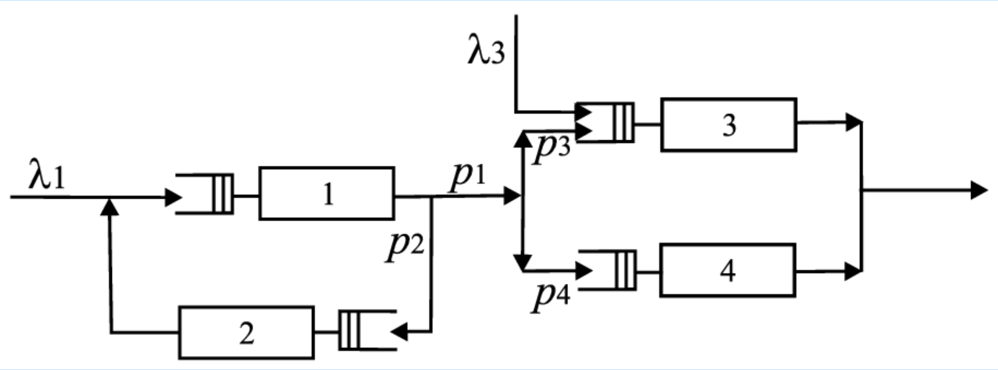
Por último, la probabilidad de que haya n_1 programas en el nodo de la *CPU* y n_2 en el nodo de *E/S* (ya sea encolados o sirviéndose) es

$$\pi(n_1, n_2) = \pi_1(n_1)\pi_2(n_2) = (1-\rho_1)\rho_1^{n_1}(1-\rho_2)\rho_2^{n_2},$$

y las medidas *L* y *W* vienen dadas por

$$L = \frac{\rho_1}{1-\rho_1} + \frac{\rho_2}{1-\rho_2} \quad \text{y} \quad W = L/\lambda.$$

Ejemplo. Sea la red de colas de la siguiente figura, con 4 procesadores.



Los trabajos llegan del exterior a los nodos 1 y 3 según un proceso de Poisson con tasas $\lambda_1=8$ y $\lambda_3=6$ trabajos por minuto, respectivamente.

Los trabajos que salen del nodo 1 pasan a procesarse en el nodo 2 con probabilidad $p_2=0.2$. Además, para aquellos trabajos que no van hacia el nodo 2, la probabilidad de ramificación hacia el nodo 3 es $p_3=0.7$.

Los tiempos de servicio en cada nodo 1, 2, 3 y 4 son exponenciales con tasas $\mu_1=14$, $\mu_2=9$, $\mu_3=17$, y $\mu_4=7$ trabajos por minuto, respectivamente. Calcular las probabilidades estacionarias y el tiempo medio de permanencia de un trabajo en la red.

Del enunciado se deduce que p_1=0.8 y p_4=0.3. Calculamos las tasas de llegadas totales a cada nodo resolviendo el sistema

$$
\begin{aligned}
\Lambda_1 &= \lambda_1 + \Lambda_2 = 8 + \Lambda_2 \\
\Lambda_2 &= p_2 \Lambda_1 = 0.2\Lambda_1 \\
\Lambda_3 &= \lambda_3 + p_1 p_3 \Lambda_1 = 6 + 0.56\Lambda_1 \\
\Lambda_4 &= p_1 p_4 \Lambda_1 = 0.24\Lambda_1
\end{aligned}
$$

cuya solución es

$$\Lambda_1 = 10, \ \Lambda_2 = 2, \ \Lambda_3 = 11.6 \ \text{y} \ \Lambda_4 = 2.4.$$

Existe distribución estacionaria porque $\Lambda_i < \mu_i$ para todo i=1,2,3,4. Por tanto,

$$
\begin{aligned}
\pi(n_1, n_2, n_3, n_4) &= \prod_{i=1}^{4} \left(\frac{\Lambda_i}{\mu_i}\right)^{n_i} \left(1 - \frac{\Lambda_i}{\mu_i}\right) \\
&= 0.046 \left(\frac{5}{7}\right)^{n_1} \left(\frac{2}{9}\right)^{n_2} \left(\frac{11.6}{17}\right)^{n_3} \left(\frac{2.4}{7}\right)^{n_4}.
\end{aligned}
$$

Calculamos el número medio de trabajos en cada nodo *i* mediante

$$L_i = \Lambda_i/(\mu_i - \Lambda_i)$$

obteniendo

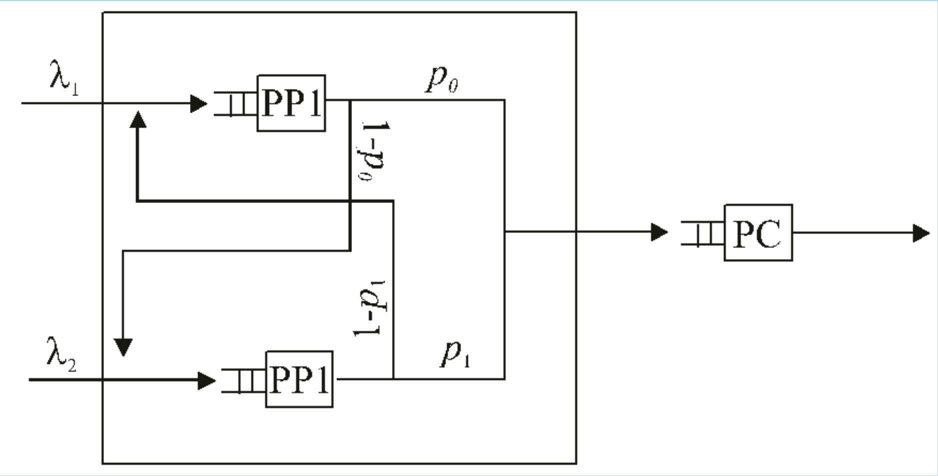
$$L_1 = 2.5,\ L_2 = 2/7,\ L_3 = 2.148 \text{ y } L_4 = 0.521 \text{ trabajos.}$$

Luego, al sumar todos, L= 5.455 trabajos. De ahí obtenemos el tiempo medio de permanencia de un trabajo en la red

$$W = \frac{L}{8+6} = 0.389 \text{ minutos.}$$

Ejercicios. Redes de Jackson abiertas

1. Consideramos la red de la figura, con dos preprocesadores (PP1 y PP2), cuya salida se encuentra conectada a un procesador central (PC):



Los trabajos llegan a los preprocesadores con tasas λ_1=2 y λ_2=4 trabajos/seg., teniendo unas tasas de servicios μ_1=10 y μ_2=20, respectivamente. Una vez que son servidos en un preprocesador pasarán al procesador Central o bien seguirán siendo preprocesados en el otro preprocesador de la red, véase la Figura. Sean p_0=0.4 y p_1=0.8. Las capacidades de las colas se consideran infinitas para los preprocesadores y finita e igual a 10 en el caso del preprocesador Central, sien-do su tasa de servicio 30 trabajos/seg.

Determinar:

a) La distribución estacionaria del número de trabajos que están siendo pre-procesados, suponiendo que los tiempos de servicio son exponenciales e independientes.

b) El número medio de trabajos que están en el sistema de preprocesamiento.

c) El tiempo medio que tarda un trabajo en ser preprocesado.

d) Indicar el tipo de cola que es el procesador Central y qué distribución de probabilidad siguen las llegadas de trabajos a él.

2. La siguiente figura representa una red abierta con 4 nodos. En cada uno de ellos se ubica un procesador